

# Protein function in the post-genomic era

David Eisenberg, Edward M. Marcotte, Ioannis Xenarios & Todd O. Yeates

Molecular Biology Institute and UCLA-DOE Laboratory of Structural Biology and Molecular Medicine, Box 951570, University of California at Los Angeles, Los Angeles, California 90095-1570, USA (e-mail: david@mbi.ucla.edu)

**Faced with the avalanche of genomic sequences and data on messenger RNA expression, biological scientists are confronting a frightening prospect: piles of information but only flakes of knowledge. How can the thousands of sequences being determined and deposited, and the thousands of expression profiles being generated by the new array methods, be synthesized into useful knowledge? What form will this knowledge take? These are questions being addressed by scientists in the field known as 'functional genomics'.**

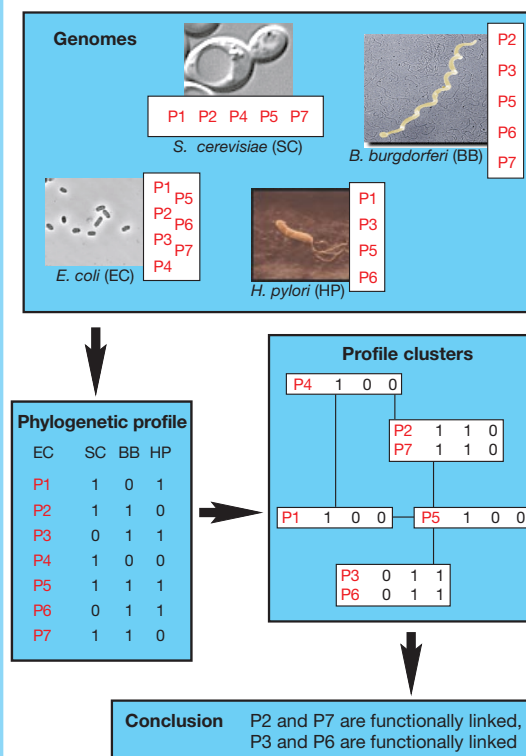
Inherent in the growing collections of genome sequences and expression profiles is knowledge about functional linkages between proteins. This knowledge can be extracted both by experimental and by computational means, as outlined below. New computational methods go beyond the traditional method of sequence homology, which seeks correlations between amino-acid sequences. Instead, correlations are sought for the inheritance of pairs of proteins into various species (for the phylogenetic profile method), for protein domains that exist both as fusions to each other and as free-standing polypeptides (for the Rosetta Stone method), or for the position of genes on chromosomes (for the gene neighbour method). Analysis of genomic and expression data by such methods produces networks of functional linkages between proteins in cells, and alters fundamentally the notion of what is meant by 'the function of a protein'.

Proteins are the main catalysts, structural elements, signalling messengers and molecular machines of biological tissues. Until recently, there have been two principal ways to learn more about the functions of protein molecules. All primary knowledge of function has come from some biochemical, genetic or structural experiment on an individual protein. But once a function has been assigned to an individual protein, one can search for other proteins with related functions by seeking proteins whose amino-acid sequences are similar to the original protein. This 'homology method' is used widely to extend knowledge of protein function from one protein to its cousins, which are presumably descended from the same common ancestral protein. The powerful BLAST programs<sup>1</sup> are used to extend experimental knowledge of protein function to new sequences in this way. By using such homology methods, roughly 40–70% of new genome sequences can be assigned to some function, the larger percentage being for well-studied prokaryotes<sup>2–4</sup>. The functional assignments by homology usually involve identification of some molecular function of the protein, but they do not place the protein in its context of cellular function, as do the methods described below.

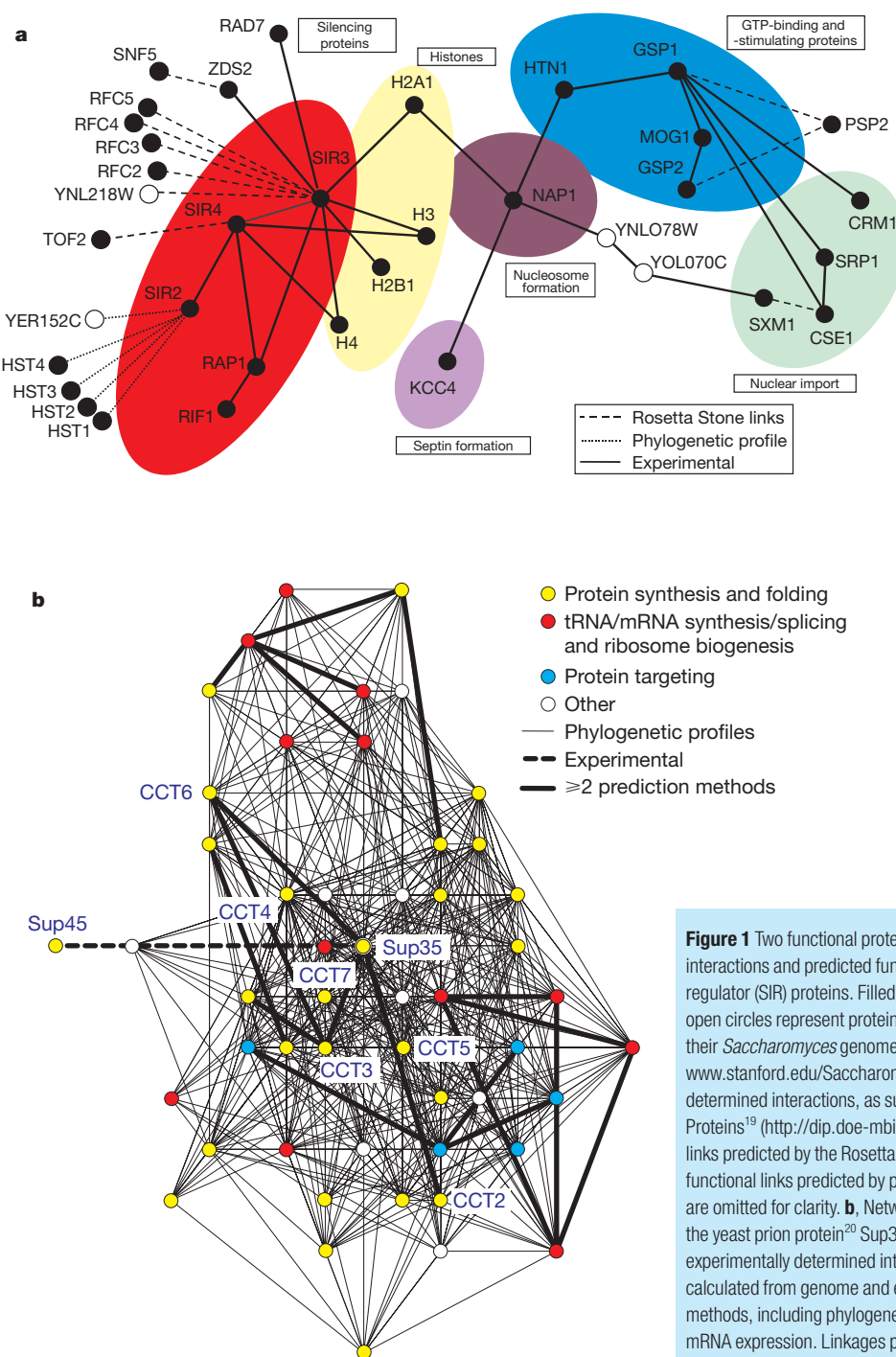
New methods have been devised to supply functional information for many proteins at once. In some cases, assignments can be made to most of the proteins encoded by the genome of an organism. These methods often detect a functional linkage between proteins. If the function of one of the proteins is known, then it can be inferred that the

## Box 1

### The method of phylogenetic profiles



The method of phylogenetic profiles is illustrated with four hypothetical genomes (top), each containing a subset of several proteins labelled P1, ..., P7. The presence or absence of each protein is indicated by 1 or 0, respectively, in the phylogenetic profiles given on the lower left. Identical profiles are clustered in boxes on the right, with profiles differing by one bit connected by lines. The conclusion at the bottom is that proteins P2 and P7 are functionally linked because they have the same phylogenetic profile and, similarly, that proteins P3 and P6 are functionally linked. Notice that two proteins that are functionally linked in this way are not in general homologues: they do not require similar sequences. This method has been described in ref. 16 with related concepts given in refs 21–25.



**Figure 1** Two functional protein networks. **a**, Network of protein interactions and predicted functional links involving silencing information regulator (SIR) proteins. Filled circles represent proteins of known function; open circles represent proteins of unknown function, represented only by their *Saccharomyces* genome sequence numbers (<http://genome-www.stanford.edu/Saccharomyces>). Solid lines show experimentally determined interactions, as summarized in the Database of Interacting Proteins<sup>19</sup> (<http://dip.doe-mbi.ucla.edu>). Dashed lines show functional links predicted by the Rosetta Stone method<sup>12</sup>. Dotted lines show functional links predicted by phylogenetic profiles<sup>16</sup>. Some predicted links are omitted for clarity. **b**, Network of predicted functional linkages involving the yeast prion protein<sup>20</sup> Sup35. The dashed line shows the only experimentally determined interaction. The other functional links were calculated from genome and expression data<sup>11</sup> by a combination of methods, including phylogenetic profiles, Rosetta stone linkages and mRNA expression. Linkages predicted by more than one method, and hence particularly reliable, are shown by heavy lines. Adapted from ref. 11.

linked proteins act in the same pathway or complex as the first protein. Even if none of the linked proteins has a known function, knowledge of the linkages is valuable in focusing future experiments and adding to the infrastructure of cellular function.

One of the most powerful of the new methods extends the two-hybrid screen to a genome-wide assay and has detected over 1,000 putative protein–protein interactions in yeast cells (see review in this issue by Pandey and Mann, pp. 837–846, and refs 5, 6). Another powerful class of methods is the analysis of correlated mRNA expression levels (see review by Lockhart and Winzler, pp. 827–836, and refs 7–9). These methods detect changes in mRNA expression in different cell types, such as a B-cell lymphoma compared with normal cells, or in

yeast cells challenged by metabolic or environmental conditions (for instance, starvation or heat). By correlating those mRNAs whose expression levels are changed, one can establish functional linkages between the proteins encoded by the correlated mRNAs<sup>10,11</sup>.

### Computational detection of functional linkages

The advent of fully sequenced genomes has facilitated the development of computational methods for establishing functional linkages between proteins. One of these computational methods is the phylogenetic profile (Box 1). A phylogenetic profile describes the pattern of presence or absence of a particular protein across a set of organisms whose genomes have been sequenced. If two proteins have

## Box 2

## The Rosetta Stone method for detecting functional linkage

## General concept


*C. elegans*

*E. coli* TrpC


The domain fusion or Rosetta Stone method for detecting functional linkage<sup>12,15</sup> is illustrated here by three examples. The top sequence in all three triplets of proteins is the fused domain or Rosetta Stone sequence; it is homologous to two separate sequences in another species. In the middle example, the genes *Pur2* and *Pur3* of yeast both encode enzymes that catalyse steps in the purine biosynthetic pathway. If it were not previously known from biochemical and genetic experiments that these enzymes are functionally linked, the linkage would be apparent from the Rosetta stone sequence Ade5,7,8 from *Caenorhabditis elegans*. Similarly, in the lower example, the fused sequence of TrpC in the *Escherichia coli* genome would inform us that the yeast proteins TrpG and TrpF are functionally linked, if we did not know already that they both catalyse steps in the biosynthesis of tryptophan.

the same phylogenetic profile (that is, the same pattern of presence or absence) in all surveyed genomes, it is inferred that the two proteins have a functional link. That is, why would two proteins always both be inherited into a new species, or neither inherited, unless the two function together? The power of the method to detect functional linkage can be appreciated when the number of possible phylogenetic profiles is considered: because each protein can be either present or absent in each genome, if there are  $n$  fully sequenced genomes, there are up to  $2^n$  phylogenetic profiles. Currently there are about 30 fully sequenced genomes in the public domain, meaning there are  $2^{30}$  ( $\sim 10^9$ ) possible phylogenetic profiles. This number far exceeds the number of protein families, so that a protein's phylogenetic profile is a nearly unique characterization of its pattern of distribution among genomes. Hence any two proteins having identical or similar phylogenetic profiles are likely to be engaged in a common pathway or complex.

Functional linkages between proteins have also been detected by analysing fusion patterns of protein domains (Box 2). Not infrequently, separate proteins A and B in one organism are expressed as a fused protein in some other species. When expressed as a fused protein, the two domains A and B are almost certainly linked in function. Thus a successful search through other genome sequences for the corresponding fused protein is powerful evidence that A and B are linked functionally. Because A and B have unrelated sequences, this type of functional linkage cannot be detected by a homology search. Also, because the fused protein has similarity to both A and B, it is termed a Rosetta Stone sequence<sup>12</sup>.

A third computational method that reveals functional linkages from genome sequences is the gene neighbour method<sup>13,14</sup>. If in several genomes the genes that encode two proteins are neighbours on the chromosome, the proteins tend to be functionally linked. This method can be powerful in uncovering functional linkages in prokaryotes, where operons are common, but also shows promise for analysing interacting proteins in eukaryotes (Box 3).

## Functional networks

When methods for detecting functional linkages are applied to all the proteins of an organism<sup>11,15</sup>, networks of interacting, functionally linked proteins can be traced out. Two examples from yeast are given in Fig. 1. Figure 1a shows interactions among histones and related proteins such as silencing proteins. These were determined mostly by experiments, but some links were predicted by the Rosetta Stone method and by phylogenetic profiles. Some of the links are to proteins known only from their genome sequences, and without other functional information; their linkage to this network indicates an intimate functional interaction among proteins involved in gene silencing, DNA packaging and nuclear transport.

Figure 1b shows a second network of functionally linked proteins

from yeast, centred on the yeast prion protein Sup35. In this network, most of the links are predicted by phylogenetic profiles, the Rosetta stone method and mRNA expression patterns. Sup35 is known to regulate translation, and it is therefore of interest that most of the predicted linkages are to other proteins involved in protein synthesis, folding and targeting. This indicates that at least some of the predicted links are meaningful. As methods improve for detecting protein linkages, it seems likely that most yeast proteins will be included in expanded versions of the networks of Fig. 1. A central feature of these networks is that most proteins interact with several other proteins.

## Validation of functional linkages

What evidence is there that functional linkages predicted by phylogenetic profiles, Rosetta stone and related methods are valid? At first glance, there is the reassurance that these methods link many proteins that are already known to function together on the basis of experiments. Examples include ribosomal proteins, proteins from the flagellar motor apparatus, and proteins in known metabolic pathways<sup>11,16</sup>. A more quantitative validation is offered by the check of 'keyword recovery'<sup>11</sup>. This simple assay compares the keyword annotations<sup>17</sup> for both members of each pair of proteins linked by one

## Box 3

## The method of correlated gene neighbours for inferring functional linkage

## Observed gene locations

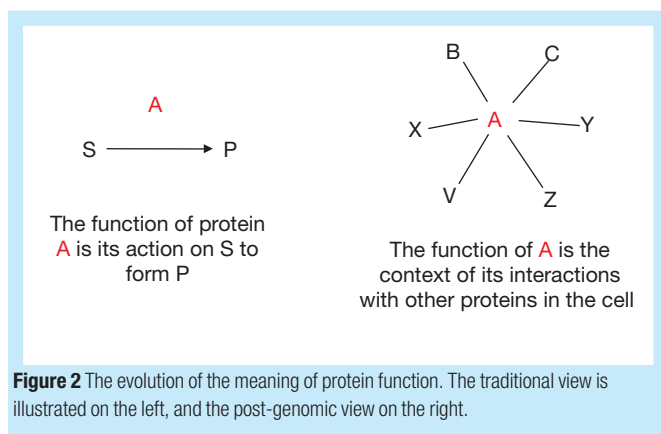


## Inferred functional linkage



If two genes (blue and yellow in the figure) are found to be neighbours in several different genomes, a functional linkage may be inferred between the proteins they encode. The method is most robust for microbial genomes but may work to some extent even for human genes where operon-like clusters are observed (see, for example, ref. 26). The gene neighbour method correctly identifies functional links among eight enzymes in the biosynthetic pathway for arginine in *Mycobacterium tuberculosis*.





of the methods. This is possible in those cases where both members of the pair have known functions. When the keywords for both members agree, there is said to be 'keyword recovery'. When keyword recovery was examined for the predicted functional linkages between yeast proteins, it was found that the individual methods showed an average signal-to-noise ratio for keyword recovery ranging between 2, for correlated mRNA expression, to 5, for the phylogenetic profiles. These values can be compared with that of 8 for direct experimental measurements of linkage. It was also found that when two of the predictive methods gave the same linkage, the signal-to-noise value was 8, the same as for direct experiments. In short, the computer-based methods for inferring function have fair reliability in general, and excellent reliability when two or more of them agree on a link.

### The post-genomic view of function

The classical view of protein function focuses on the action of a single protein molecule. This action may be the catalysis of a given reaction or the binding of a small or large molecule. Today this local function is sometimes termed the 'molecular function' of the protein to distinguish it from an expanded view of function (Fig. 2). In the expanded view of protein function, a protein is defined as an element in the network of its interactions. Various terms have been coined for this expanded notion of function, such as 'contextual function' or 'cellular function' (see, for example, ref. 18). Whatever the term, the idea is that each protein in living matter functions as part of an extended web of interacting molecules.

In conclusion, the availability of fully sequenced genomes and the enormous amount of data on the co-expression of mRNAs opens new ways to analyse protein function. The new methods establish functional links between pairs of proteins, and interconnecting links form networks of functionally interacting proteins. Some of the functional

linkages reflect metabolic or signalling pathways; other linkages reflect the formation of complexes of macromolecules such as ribosomes. Often it is possible to understand the cellular functions of uncharacterized proteins through their linkages to characterized proteins. In broader terms, the networks of linkages offer a new view of the meaning of protein function, and in time should offer a deepened understanding of the functioning of cells. □

1. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
2. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
3. Chervitz, S. A. *et al.* Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* **282**, 2022–2028 (1998).
4. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36 (2000).
5. Uetz, P. *et al.* A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
6. Ito, T. *et al.* Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl Acad. Sci. USA* **97**, 1143–1147 (2000).
7. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
8. Lashkari, D. A. *et al.* Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl Acad. Sci. USA* **94**, 13057–13062 (1997).
9. Brown, P. O. & Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nature Genet.* **21**, 33–37 (1999).
10. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
11. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83–86 (1999).
12. Marcotte, E. M. *et al.* Detecting protein function and protein–protein interactions from genome sequences. *Science* **285**, 751–753 (1999).
13. Dandekar, T., Snel, B., Huynen, M. & Bork, P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23**, 324–328 (1998).
14. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA* **96**, 2896–2901 (1999).
15. Enright, A. J., Iliopoulos, I., Kyripides, N. C. & Ouzounis, C. A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86–90 (1999).
16. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA* **96**, 4285–4288 (1999).
17. Andrade, M. A. & Valencia, A. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics* **14**, 600–607 (1998).
18. Kim, S. H. Structural genomics of microbes: an objective. *Curr. Opin. Struct. Biol.* (in the press).
19. Xenarios, I. *et al.* DIP: the Database of Interacting Proteins. *Nucleic Acids Res.* **28**, 289–291 (2000).
20. Wickner, R. B. [URE3] as an altered URE2 protein: evidence for a prion analog in *Saccharomyces cerevisiae*. *Science* **264**, 566–569 (1994).
21. Bork, P. *et al.* Predicting function: from genes to genomes and back. *J. Mol. Biol.* **283**, 707–725 (1998).
22. Huynen, M., Dandekar, T. & Bork, P. Differential genome analysis applied to the species-specific features of *Helicobacter pylori*. *FEBS Lett.* **426**, 1–5 (1998).
23. Hegyi, H. & Gerstein, M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147–164 (1999).
24. Ouzounis, C. & Kyripides, N. The emergence of major cellular processes in evolution. *FEBS Lett.* **390**, 119–123 (1996).
25. Gaasterland, T. & Ragan, M. A. Constructing multigenome views of whole microbial genomes. *Microb. Comp. Genomics* **3**, 177–192 (1998).
26. Wu, Q. & Maniatis, T. A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell* **97**, 779–790 (1999).